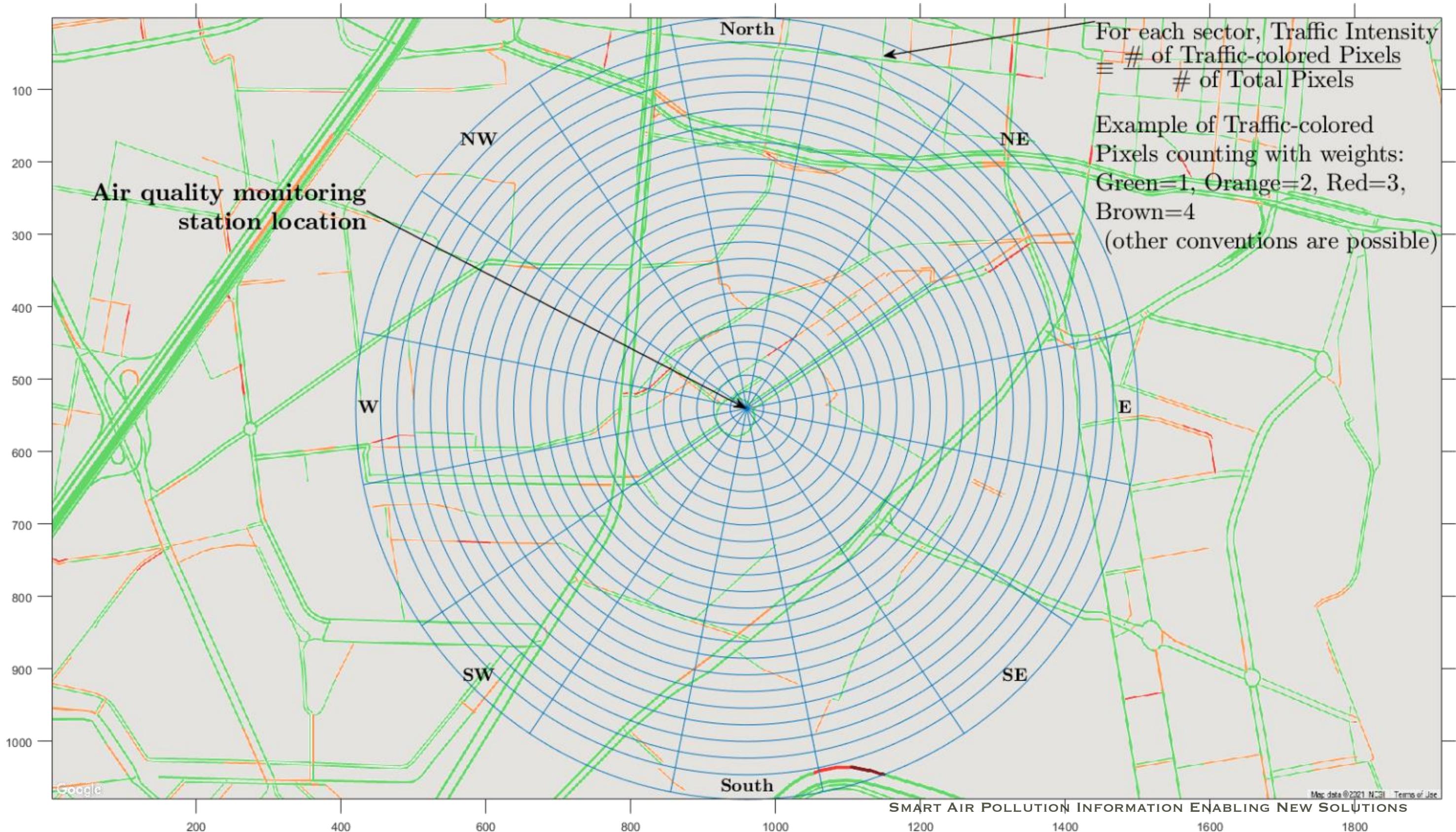
Traffic Intensity Model

Google Maps images with traffic layer/colouring: green, orange, red, dark red
% of coloured pixels in annular sectors -> traffic intensity; physical density "field"

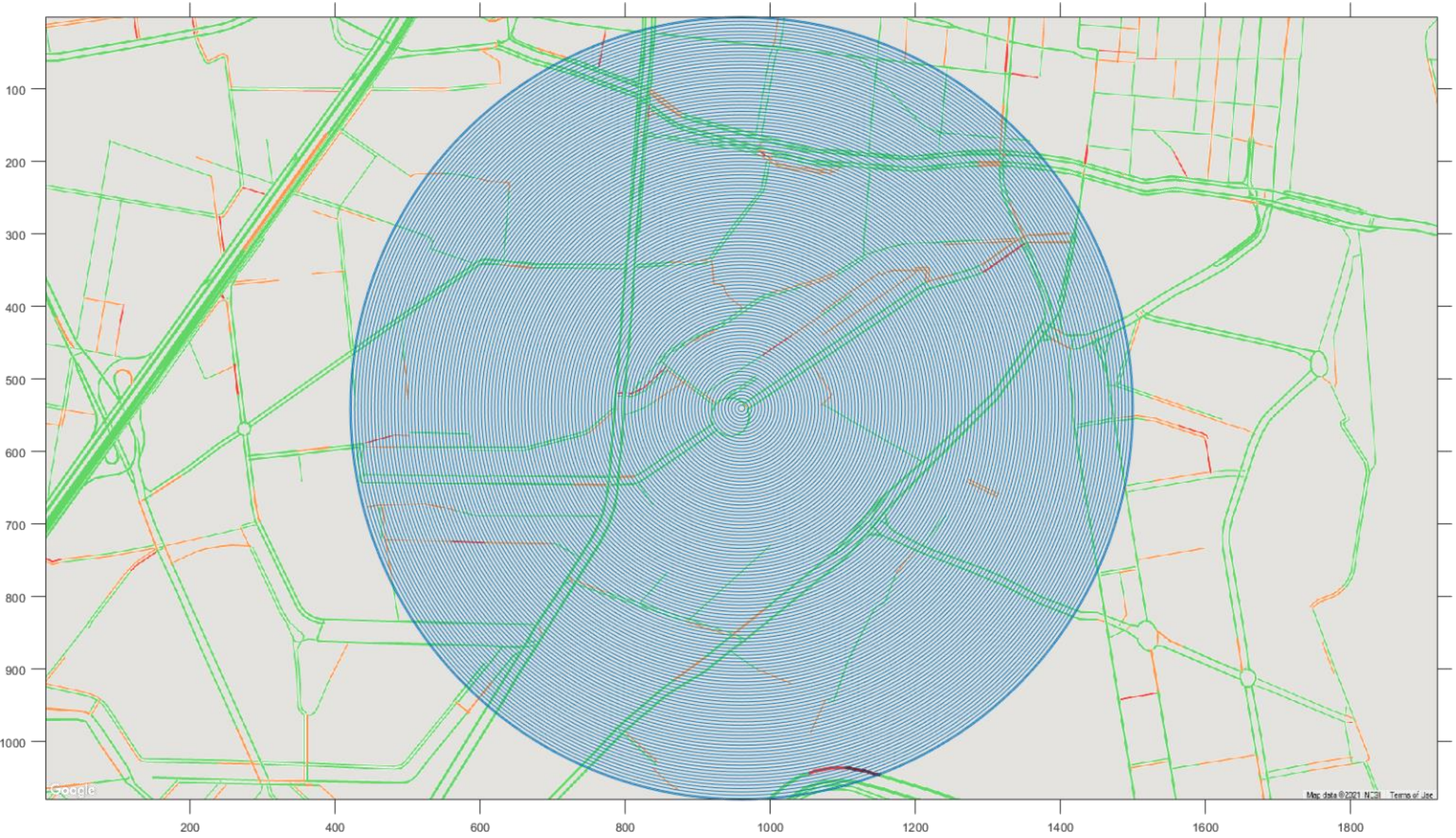
Dimension reduction
from HD image 1920x1080:

(4 colours)x(16 angles)x(118 rings)
Or aggregating the angles:
(4 colours)x(118 rings)

Only 23 rings in this plot
118 rings shown in the next page



Traffic Intensity Model



Width of 118 concentric rings: 10m

Regression Modeling: data matrices

Spatio-Temporal Data Matrices X (traffic predictors; centered as X_0 with 0 mean & rescaled with Var=1) and Y (pollutants responses; centered & rescaled as Y_0):

$$
\begin{array}{c}
\text{observations (temporal)} \\ \downarrow
\end{array}
\;
\overset{\text{predictor variables (spatial)}\;\longrightarrow}{
\begin{array}{cc}
 & \begin{array}{ccc}\text{1st predictor} & \cdots & p\text{-th predictor}\end{array} \\
\begin{array}{c}\text{1st obs.} \\ \vdots \\ n\text{-th obs.}\end{array} &
\begin{bmatrix}
\vert & \text{---} & X_0(1,:) & \text{---} & \vert \\
\vert & \ddots & \vdots & & \vert \\
X_0(:,1) & \cdots & X\,(\text{centered as } X_0) & \cdots & X_0(:,p) \\
\vert & & \vdots & \ddots & \vert \\
\vert & \text{---} & X_0(n,:) & \text{---} & \vert
\end{bmatrix}
\end{array}}
\;
\overset{\text{response variables (spatial)}\;\longrightarrow}{
\begin{array}{c}
\begin{array}{ccc}\text{CO} & \cdots & \text{NO2}\end{array} \\
\begin{bmatrix}
\vert & \text{---} & Y_0(1,:) & \text{---} & \vert \\
\vert & \ddots & \vdots & & \vert \\
Y_0(:,1) & \cdots & Y\,(\text{centered as } Y_0) & \cdots & Y_0(:,m=9) \\
\vert & & \vdots & \ddots & \vert \\
\vert & \text{---} & Y_0(n,:) & \text{---} & \vert
\end{bmatrix}
\end{array}}
$$

where $p = $ (4 colors)×(118 rings) (or (4 colors)×(16 angles)×(118 rings)) traffic predictor variables, $m = 9$ pollutant response variables;

n observations, depending on station/sensor (each has different # of missing/null readings)

Regression Modeling: PLSR

Purpose of our modeling is two-fold:

(1) (interpretations) get traffic activities $\xrightarrow{\text{mapping}}$ pollutant concentrations, and reveal insight on their detailed relations, and

(2) (predictions) predict pollutant concentrations based on traffic activities.

Most black-box machine learning techniques are good at (2) only;

Partial least squares (PLS) regression formulation:

$$\begin{cases} X_0 = X_S X_L^{\mathsf{T}} + X_{\text{residuals}} \\ Y_0 = Y_S Y_L^{\mathsf{T}} + Y_{\text{residuals}} \end{cases}$$

Spatio-Temporal Data

- Example of spatio-temporal data: sequence of images, videos, etc.
- Often centered to have 0-mean and rescaled to have unit variance

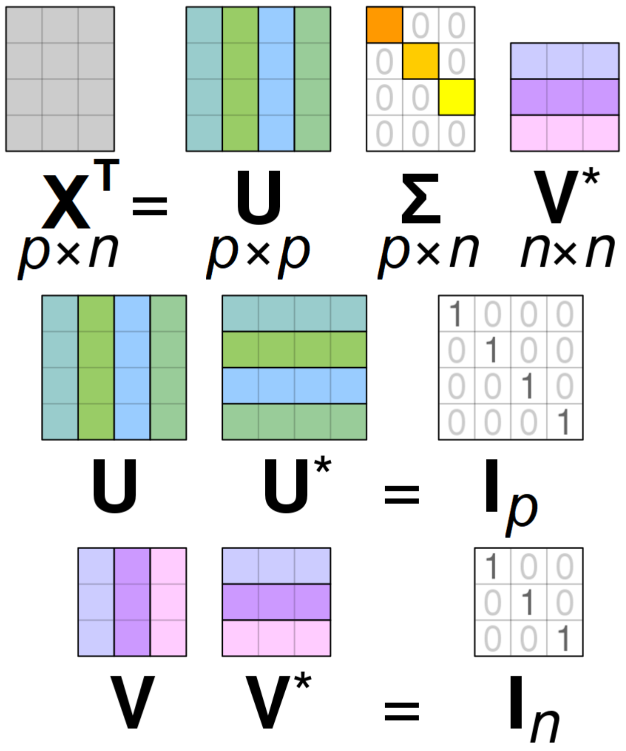
$$\xrightarrow{\text{variables (spatial)}}$$

$$\text{observations (temporal)} \downarrow \quad \begin{array}{cc} & \begin{array}{ccccc} \text{1st variable} & & \cdots & & p\text{-th variable} \end{array} \\ \begin{array}{c} \text{1st obs.} \\ \\ \vdots \\ \\ n\text{-th obs.} \end{array} & \left[\begin{array}{ccccc} | & \text{—} & X_0(1,:) & \text{—} & | \\ | & \ddots & \vdots & & | \\ X_0(:,1) & \cdots & X \text{ (centered as } X_0) & \cdots & X_0(:,p) \\ | & & \vdots & \ddots & | \\ | & \text{—} & X_0(n,:) & \text{—} & | \end{array}\right] \end{array}$$

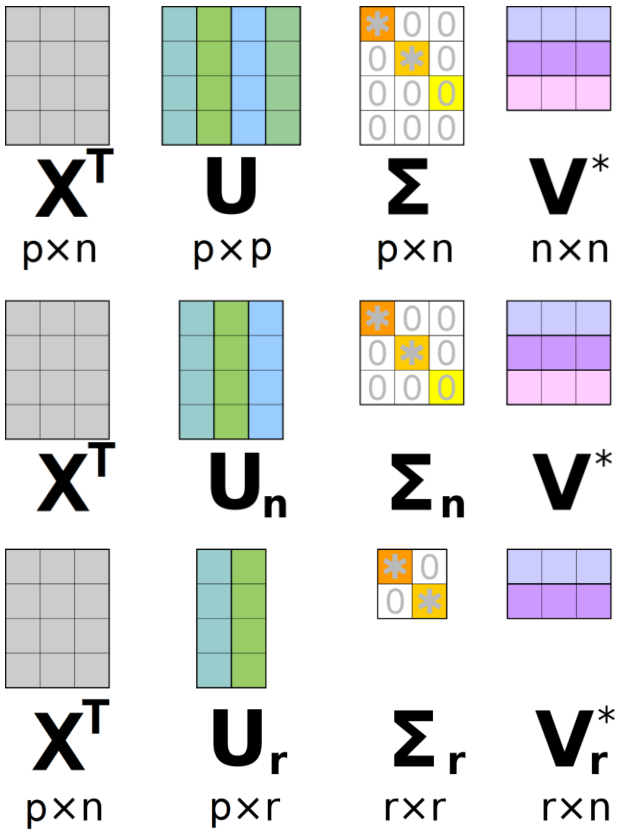
- Problem: both n and p are very large, variables may be correlated;
- Fitting a naïve model using all p variables $\Rightarrow$ overfitting, and too “black-box”
- How to effectively reduce p without losing too much information, and even extract or reveal some useful insight or interpretation?

Preliminary: Singular Value Decomposition (SVD)

- Singular value decomposition: $X^{\mathrm{T}} = U\Sigma V^*$
- For real matrices, $U^* = U^{\mathrm{T}}, V^* = V^{\mathrm{T}}$, and
- U & V are both orthogonal matrices:

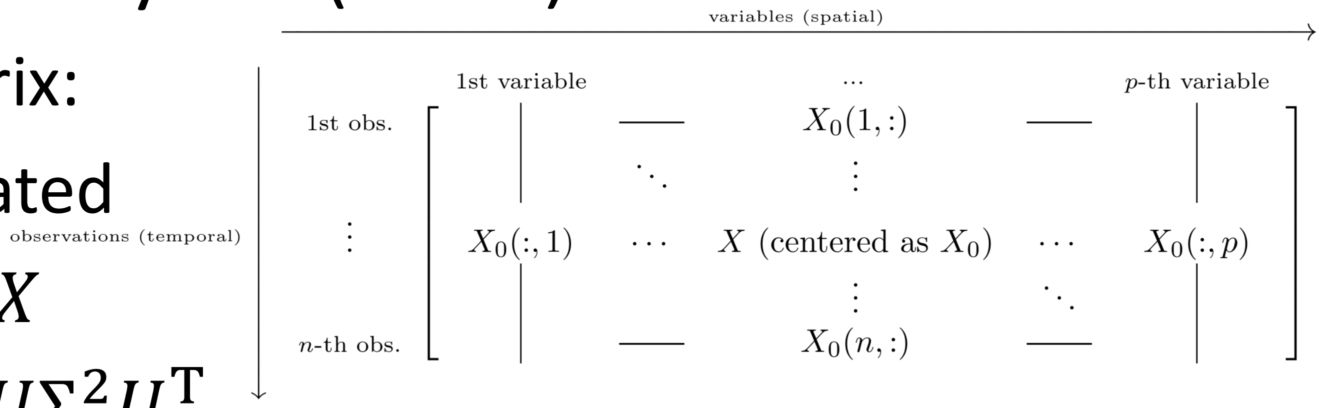


- p (spatial; variables) $> n$ (temporal; observations), reduced SVD:
- Can reduce further:



- Diagonal line of Σ: non-increasing, truncate the diagonal Σ to 1st r values, and the columns of U and V
- Columns of U are still orthonormal basis; same for V

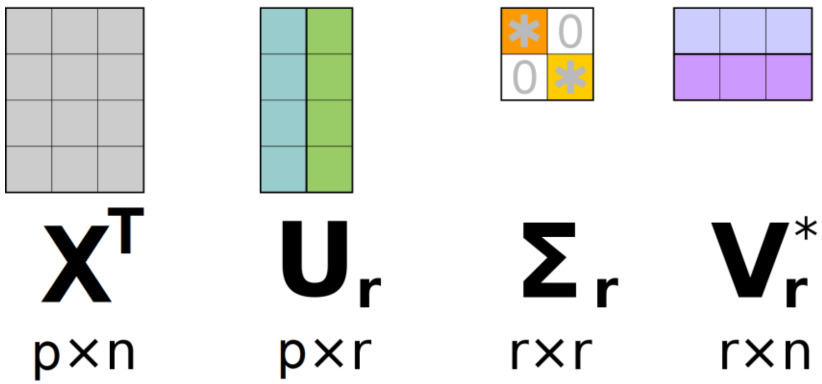
Principal Component Analysis (PCA)



- Recall the spatio-temporal data matrix:
- Variables (columns) are often correlated
- Covariance matrix of variables $\propto X^{\mathrm{T}}X$
- Using SVD: $X^{\mathrm{T}}X = U\Sigma V^{\mathrm{T}}V\Sigma^{\mathrm{T}}U^{\mathrm{T}} = U\Sigma^2U^{\mathrm{T}}$
- $\Sigma^2 \propto$ eigenvalues of covariance matrix of variables (columns in X)
- Columns in U are the corresponding eigenvectors, called Principal Components
- $X^{\mathrm{T}} = U\Sigma V^{\mathrm{T}}$, so $U^{\mathrm{T}}X^{\mathrm{T}} = \Sigma V^{\mathrm{T}}$, i.e., projecting each observation/data point to columns of U, and get ΣV^{T} of $r \times n$ (truncated to r components)
- Rows of $\Sigma V^{\mathrm{T}} = X_S^{\mathrm{T}}$ called “scores”,

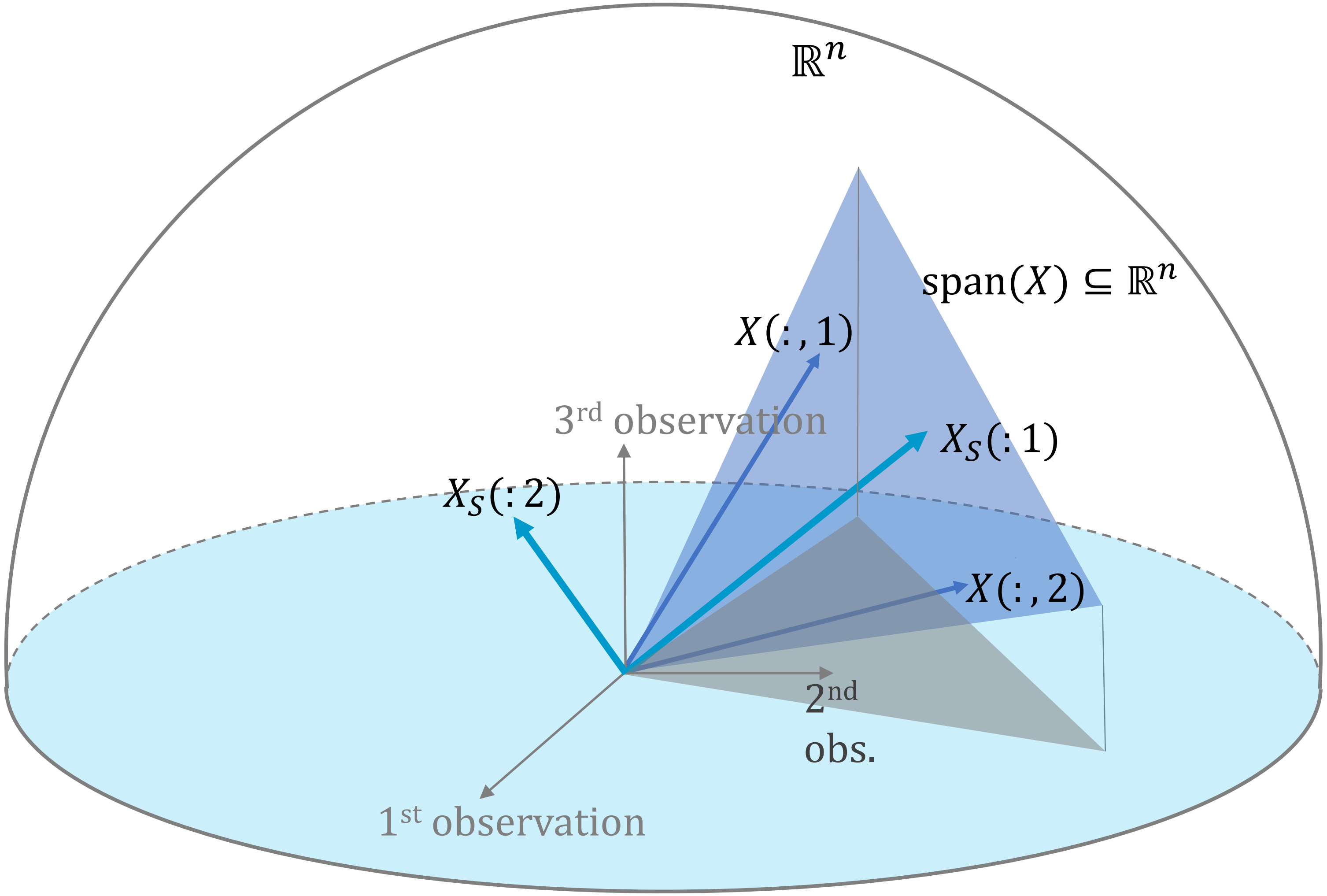
capturing major variances up to r elements of Σ^2

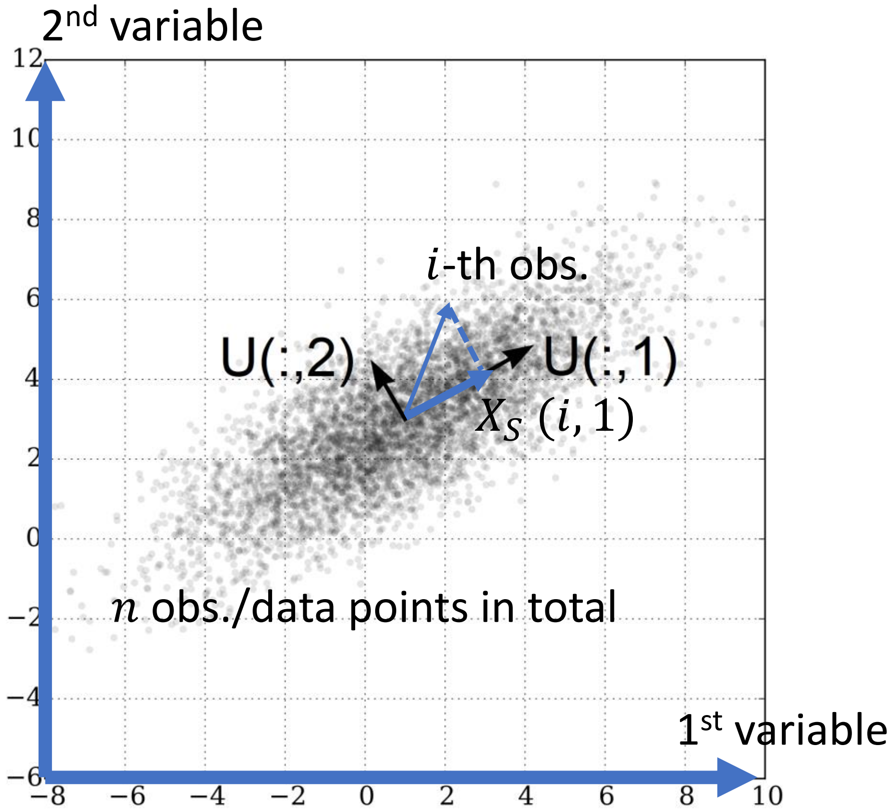
- $\|X_S(:,i)\|^2 = \Sigma^2(i,i)$



Principal Component Analysis (PCA)

- Geometric interpretation:





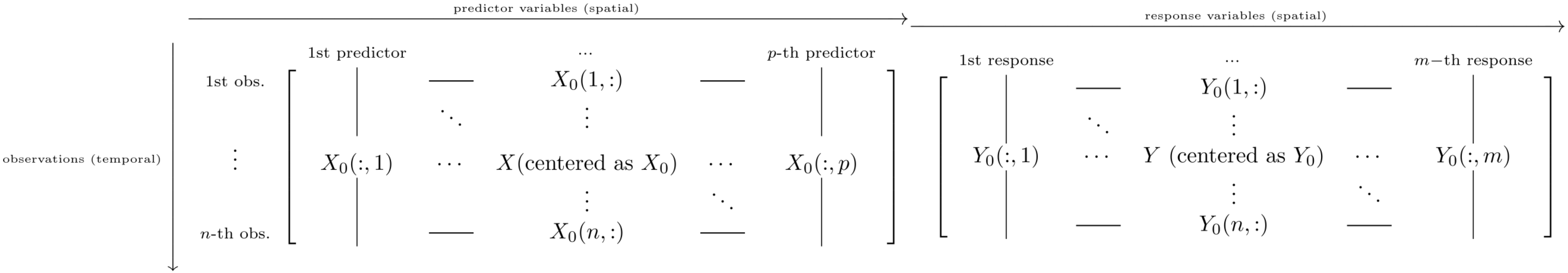
$$\|X_S(:,i)\|^2 = \Sigma^2(i,i)$$

Non-increasing along diagonal of Σ

Truncate up to $X_S(:,r)$ to reduce dimension

Partial Lease Squares (PLS)

- PCA only considers 1 spatio-temporal data matrix
- In regression analysis, we have both predictor variables (predictors, independent variables; x) and response variables (responses, dependent variables; y), organized in X (centered as X_0 with 0 mean & rescaled with Var=1) and Y (centered & rescaled as Y_0):



Partial least squares (PLS) regression formulation:

$$\begin{cases} X_0 = X_S X_L^{\mathsf{T}} + X_{\text{residuals}} \\ Y_0 = Y_S Y_L^{\mathsf{T}} + Y_{\text{residuals}} \end{cases}$$

Regression Modeling: PLSR

$$\begin{cases} X_0 = X_S X_L^{\mathsf{T}} + X_{\text{residuals}} \\ Y_0 = Y_S Y_L^{\mathsf{T}} + Y_{\text{residuals}} \end{cases}$$

predictor variables (spatial)

observations (temporal)

$$\begin{array}{cc}
 & \begin{array}{ccccc} \text{1st predictor} & & \cdots & & p\text{-th predictor} \end{array} \\
\begin{array}{c} \text{1st obs.} \\ \vdots \\ n\text{-th obs.} \end{array} &
\left[\begin{array}{ccccc}
| & \text{---} & X_0(1,:) & \text{---} & | \\
| & \ddots & \vdots & & | \\
X_0(:,1) & \cdots & X\ (\text{centered as } X_0) & \cdots & X_0(:,p) \\
| & & \vdots & \ddots & | \\
| & \text{---} & X_0(n,:) & \text{---} & |
\end{array}\right]
\end{array} =$$

PLS components/modes number

observations (temporal)

$$\begin{array}{cc}
 & \begin{array}{ccccc} \text{1st PLS comp.} & & \cdots & & n_{\text{comp}}\text{-th PLS comp.} \end{array} \\
\begin{array}{c} \text{1st obs.} \\ \vdots \\ n\text{-th obs.} \end{array} &
\left[\begin{array}{ccccc}
| & \text{---} & X_S(1,:) & \text{---} & | \\
| & \ddots & \vdots & & | \\
X_S(:,1) & \cdots & X_S\ (X\ \text{scores}) & \cdots & X_S(:,n_{\text{comp}}) \\
| & & \vdots & \ddots & | \\
| & \text{---} & X_S(n,:) & \text{---} & |
\end{array}\right]
\end{array}$$

predictor variables (spatial)

$$\begin{array}{cc}
 & \begin{array}{ccccc} \text{1st predictor} & & \cdots & & p\text{-th predictor} \end{array} \\
\begin{array}{c} \text{1st PLS comp.} \\ \vdots \\ n_{\text{comp}}\text{-th comp.} \end{array} &
\left[\begin{array}{ccccc}
| & \text{---} & X_L^{\mathrm{T}}(1,:) & \text{---} & | \\
| & \ddots & \vdots & & | \\
X_L^{\mathrm{T}}(:,1) & \cdots & X_L^{\mathrm{T}}\ (X\ \text{loading}) & \cdots & X_L^{\mathrm{T}}(:,p) \\
| & & \vdots & \ddots & | \\
| & \text{---} & X_L^{\mathrm{T}}(n_{\text{comp}},:) & \text{---} & |
\end{array}\right]
\end{array}$$

response variables (spatial)

observations (temporal)

$$\begin{array}{cc}
 & \begin{array}{ccccc} \text{CO} & & \cdots & & \text{NO2} \end{array} \\
\begin{array}{c} \text{1st obs.} \\ \vdots \\ n\text{-th obs.} \end{array} &
\left[\begin{array}{ccccc}
| & \text{---} & Y_0(1,:) & \text{---} & | \\
| & \ddots & \vdots & & | \\
Y_0(:,1) & \cdots & Y\ (\text{centered as } Y_0) & \cdots & Y_0(:,p) \\
| & & \vdots & \ddots & | \\
| & \text{---} & Y_0(n,:) & \text{---} & |
\end{array}\right]
\end{array} =$$

PLS components/modes number

observations (temporal)

$$\begin{array}{cc}
 & \begin{array}{ccccc} \text{1st PLS comp.} & & \cdots & & n_{\text{comp}}\text{-th PLS comp.} \end{array} \\
\begin{array}{c} \text{1st obs.} \\ \vdots \\ n\text{-th obs.} \end{array} &
\left[\begin{array}{ccccc}
| & \text{---} & Y_S(1,:) & \text{---} & | \\
| & \ddots & \vdots & & | \\
Y_S(:,1) & \cdots & Y_S\ (Y\ \text{scores}) & \cdots & Y_S(:,n_{\text{comp}}) \\
| & & \vdots & \ddots & | \\
| & \text{---} & Y_S(n,:) & \text{---} & |
\end{array}\right]
\end{array}$$

response variables (spatial)

$$\begin{array}{cc}
 & \begin{array}{ccccc} \text{CO} & & \cdots & & \text{NO2} \end{array} \\
\begin{array}{c} \text{1st PLS comp.} \\ \vdots \\ n_{\text{comp}}\text{-th comp.} \end{array} &
\left[\begin{array}{ccccc}
| & \text{---} & Y_L^{\mathrm{T}}(1,:) & \text{---} & | \\
| & \ddots & \vdots & & | \\
Y_L^{\mathrm{T}}(:,1) & \cdots & Y_L^{\mathrm{T}}\ (Y\ \text{loading}) & \cdots & Y_L^{\mathrm{T}}(:,m) \\
| & & \vdots & \ddots & | \\
| & \text{---} & Y_L^{\mathrm{T}}(n_{\text{comp}},:) & \text{---} & |
\end{array}\right]
\end{array}$$

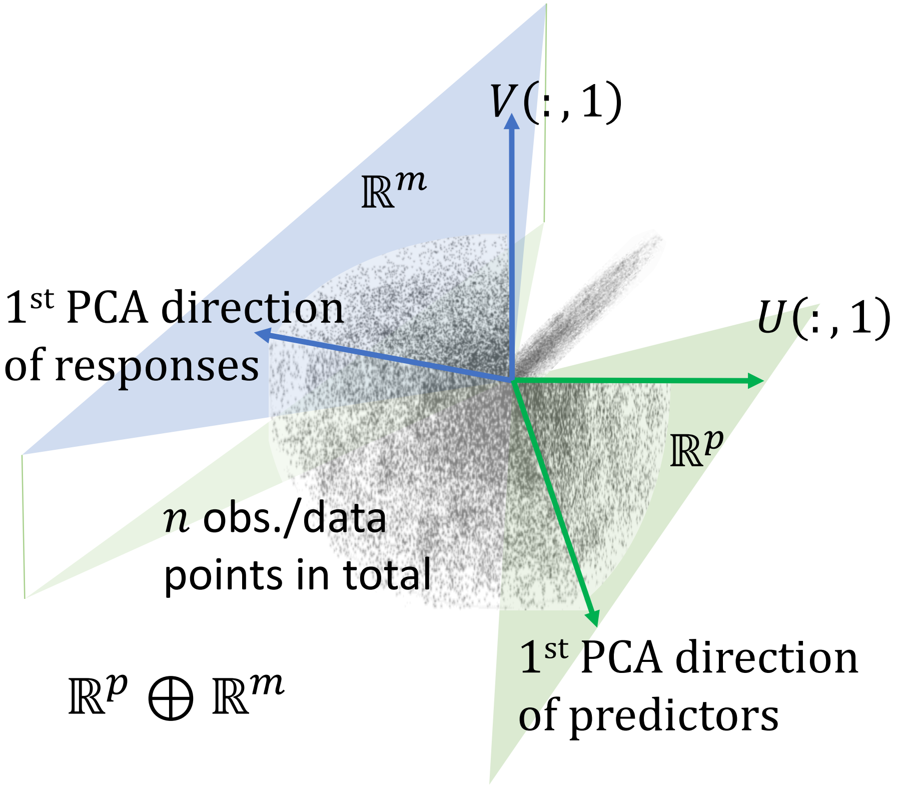
Partial Lease Squares (PLS)

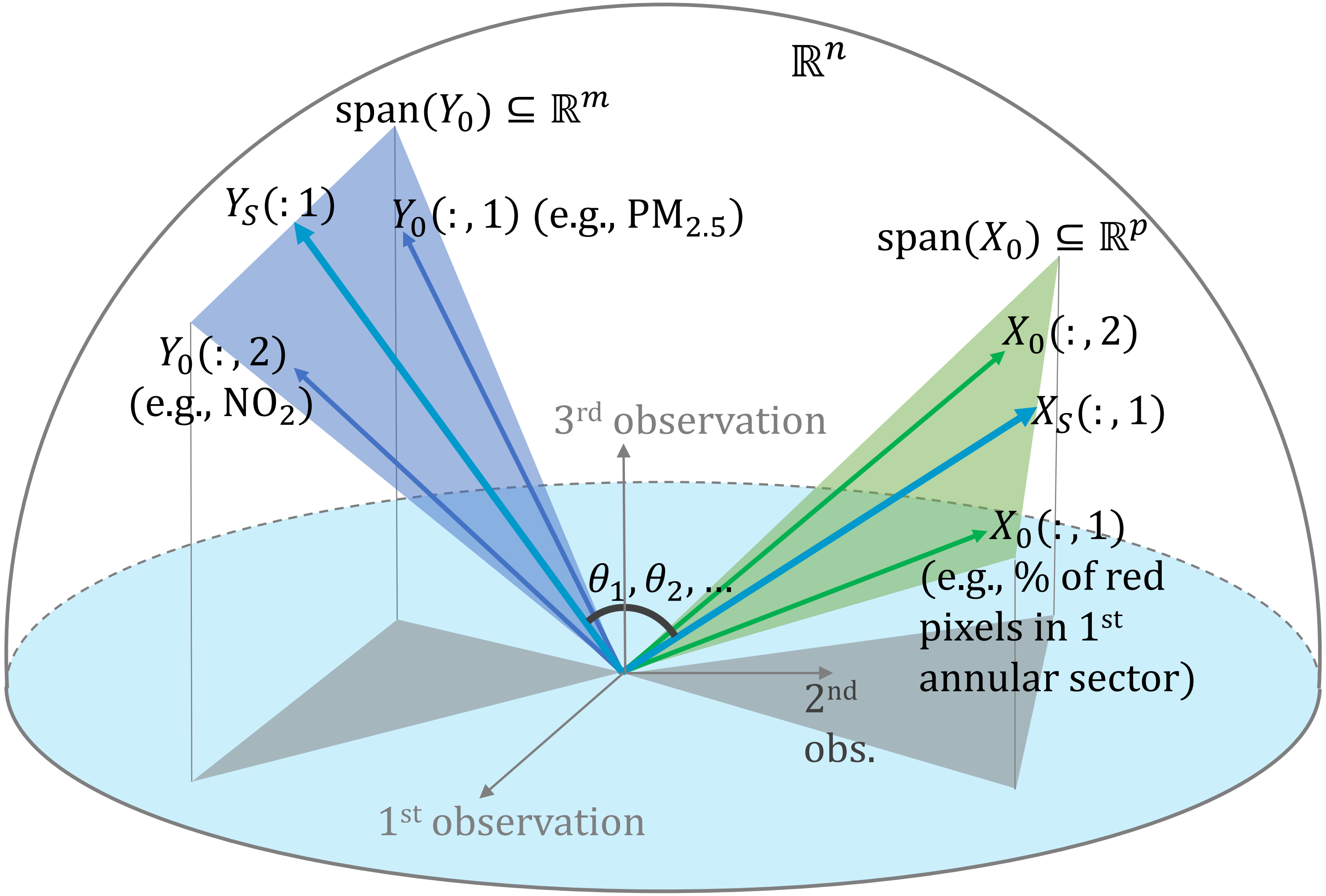
- Partial least squares (PLS) regression formulation:

$$\begin{cases} X_0 = X_S X_L^{\top} + X_{\text{residuals}} \\ Y_0 = Y_S Y_L^{\top} + Y_{\text{residuals}} \end{cases}$$

- It aims to find correlation between X_0 and Y_0
- Recall SVD and covariance matrix (p-by-m): $X_0^{\top} Y_0 = U \Sigma V^{\top}$
- Again, columns of U and V are orthonormal basis: $U^{\top} X_0^{\top} Y_0 V = X_S^{\top} Y_S = \Sigma$
- Projection of X_0 on columns of U obtains predictor scores X_S;
- Projection of Y_0 on columns of V obtains response scores Y_S;
- $X_S^{\top} Y_S$ is inner products between columns of X_S and Y_S
- If X_0 and Y_0 are centered and rescaled/normalized, projected on to U and V to get X_S and Y_S, so $X_S^{\top} Y_S = \Sigma$ contains cosines between columns of X_S and Y_S.
- Σ is diagonal, non-increasing order: cosine (and hence correlation) between $X_S(:,1)$ and $Y_S(:,1)$ is maximal; angle is minimal

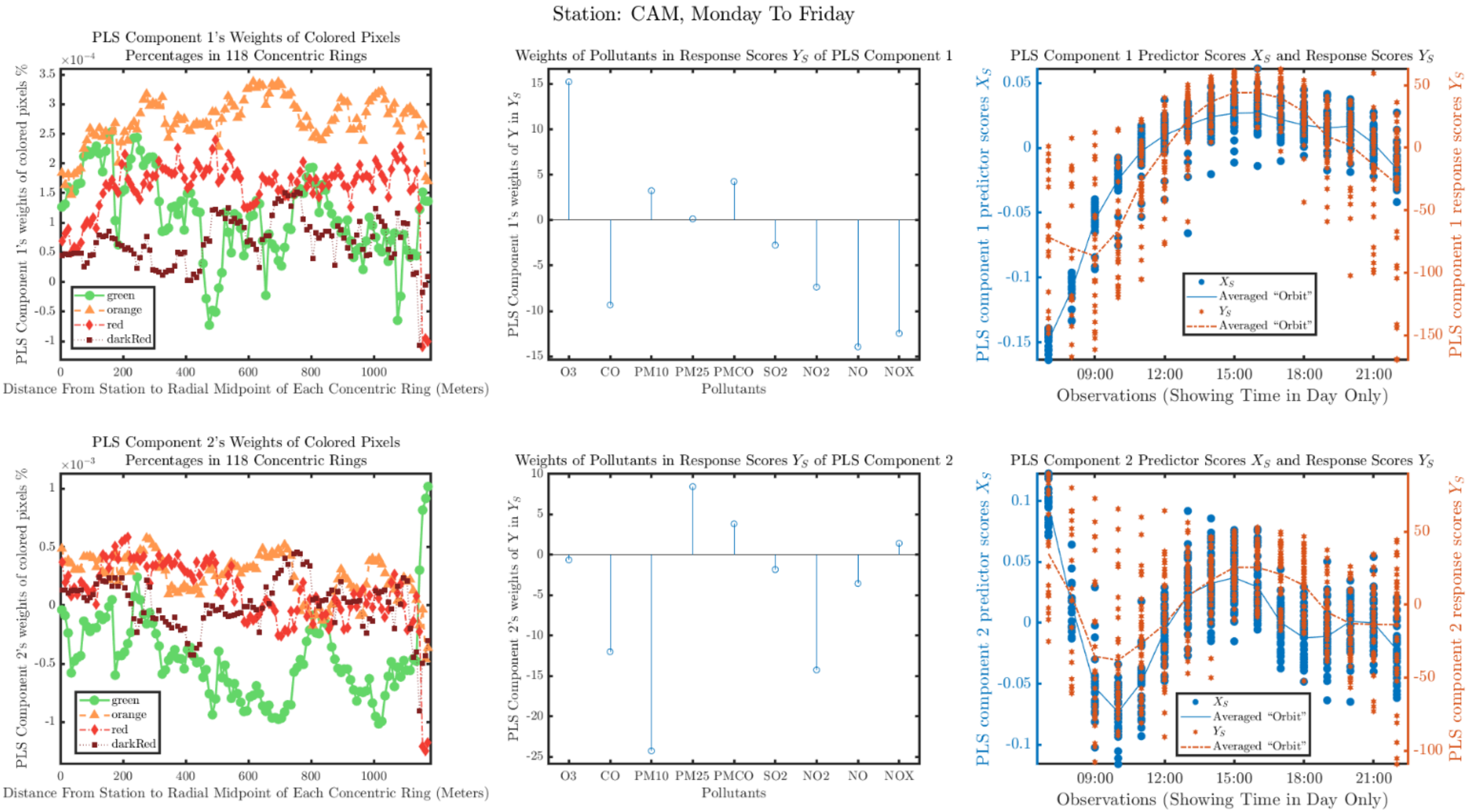
Geometric Interpretation: principal angles between flats (subspaces)



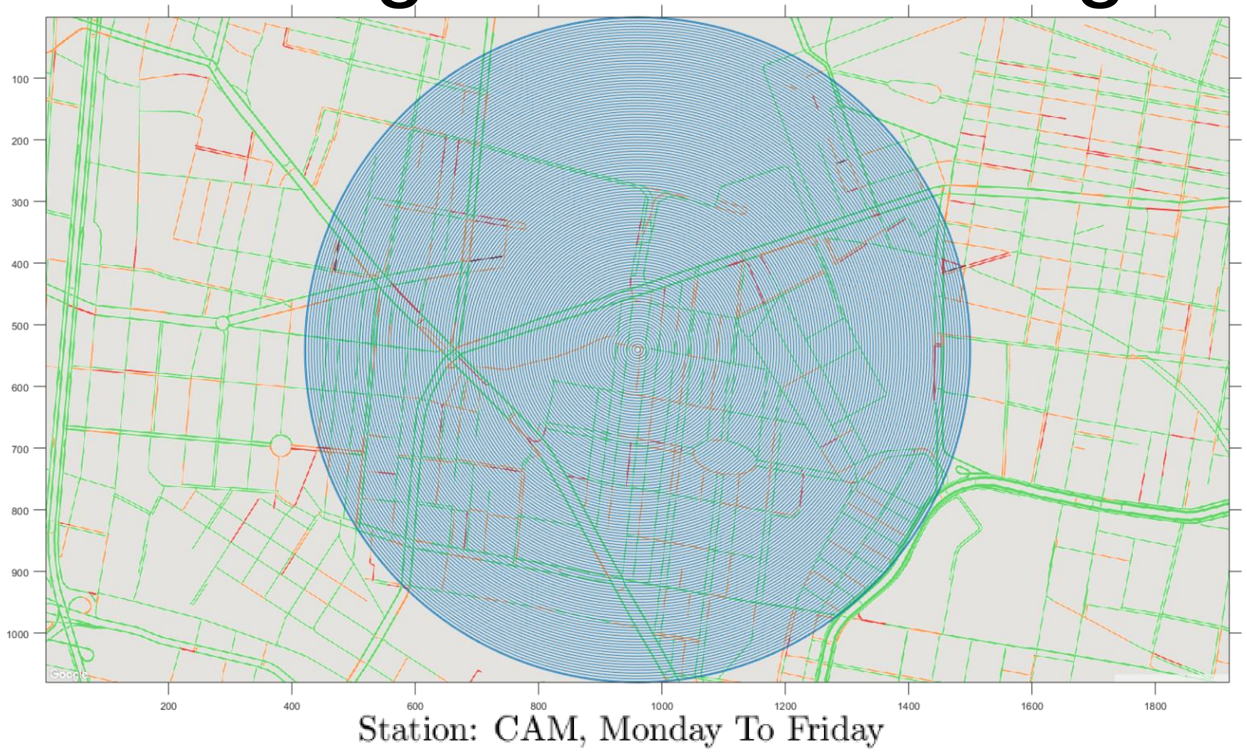


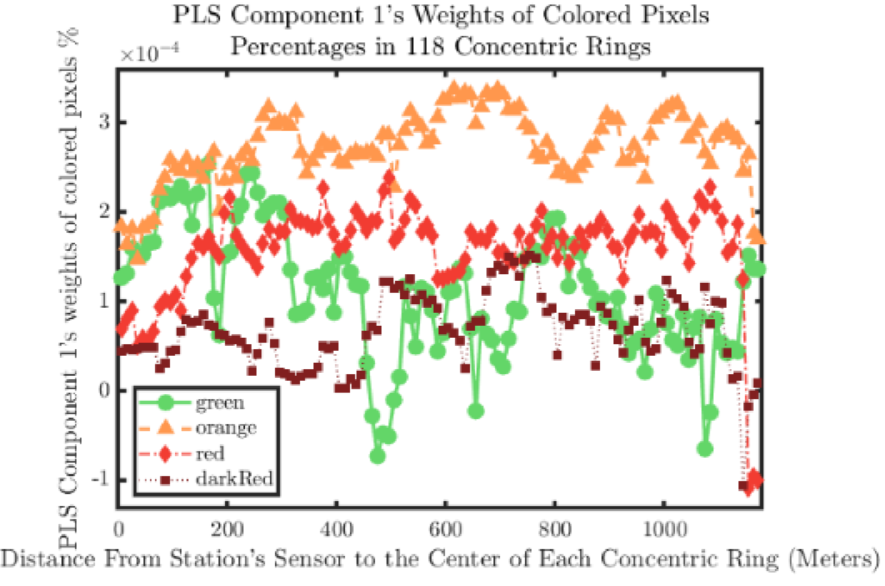
Interpretations and Insight from PLS Regression Modeling

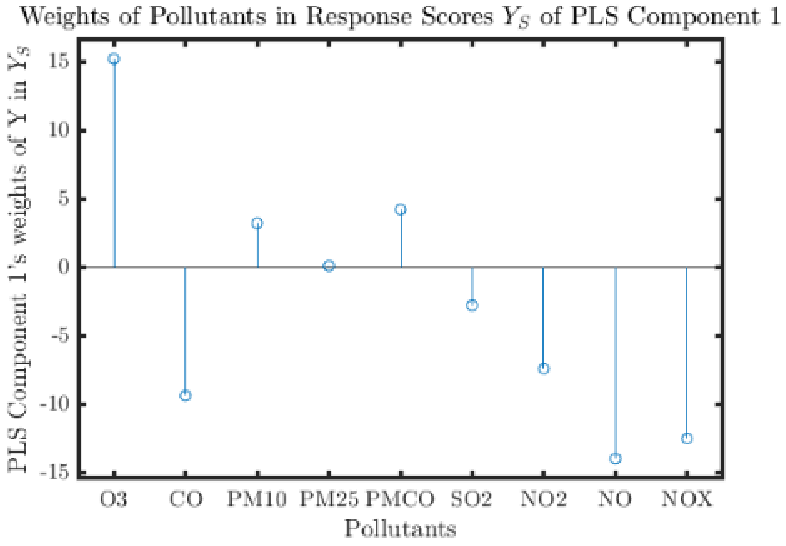
1st PLS component has physically meaningful interpretation:

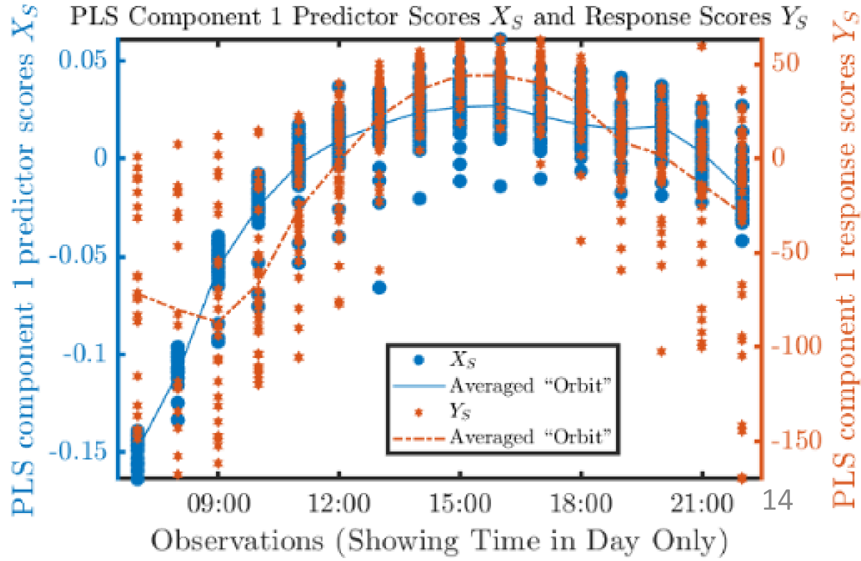


Interpretations and Insight from PLS Regression Modeling

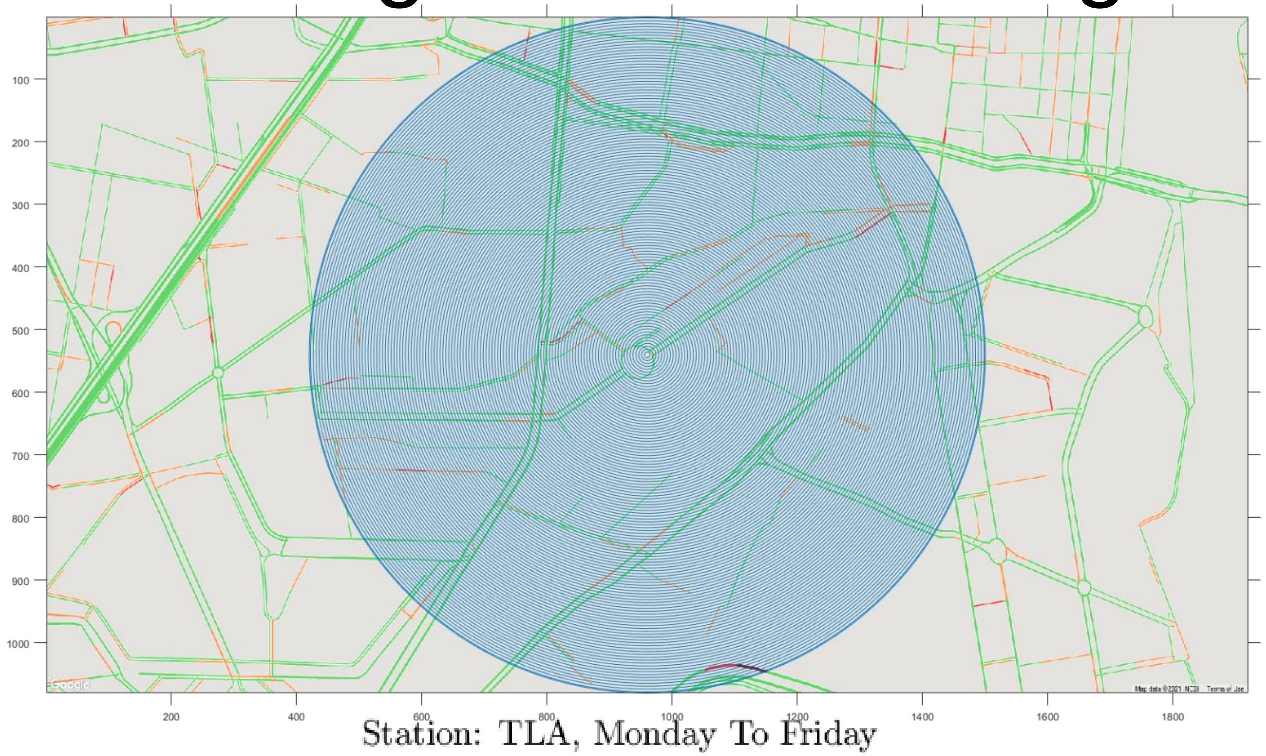


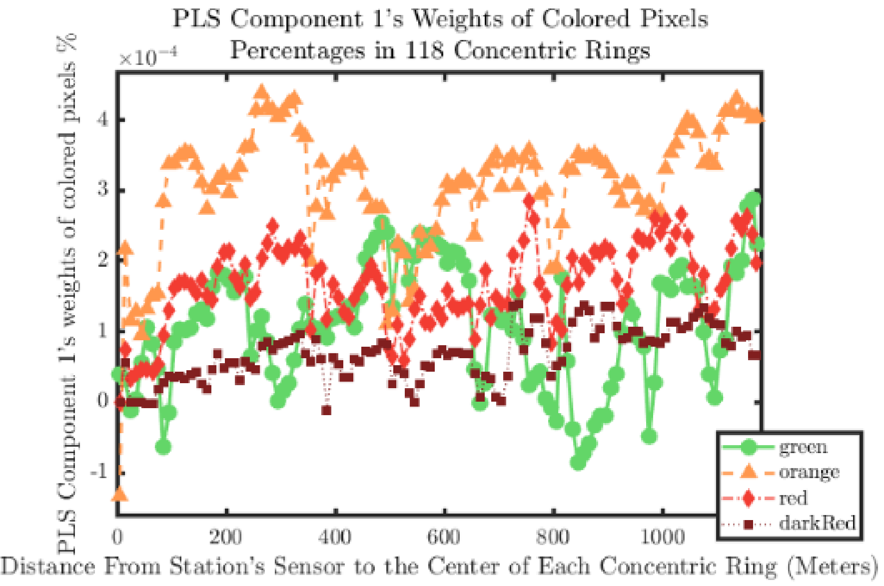


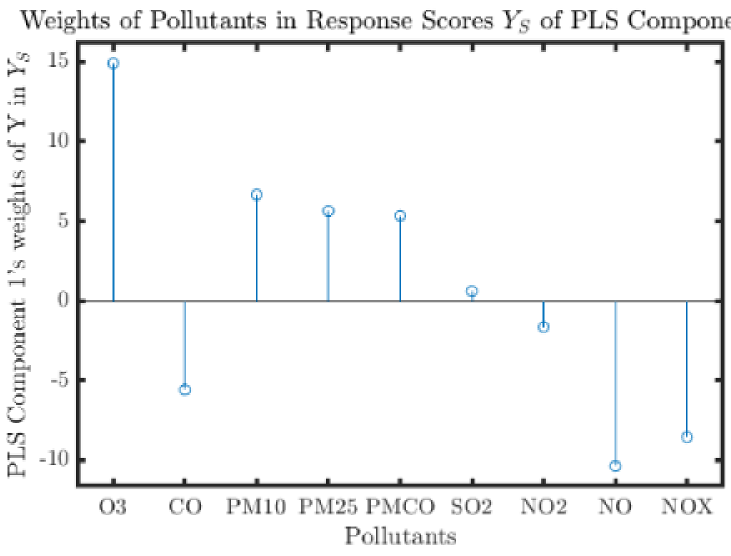


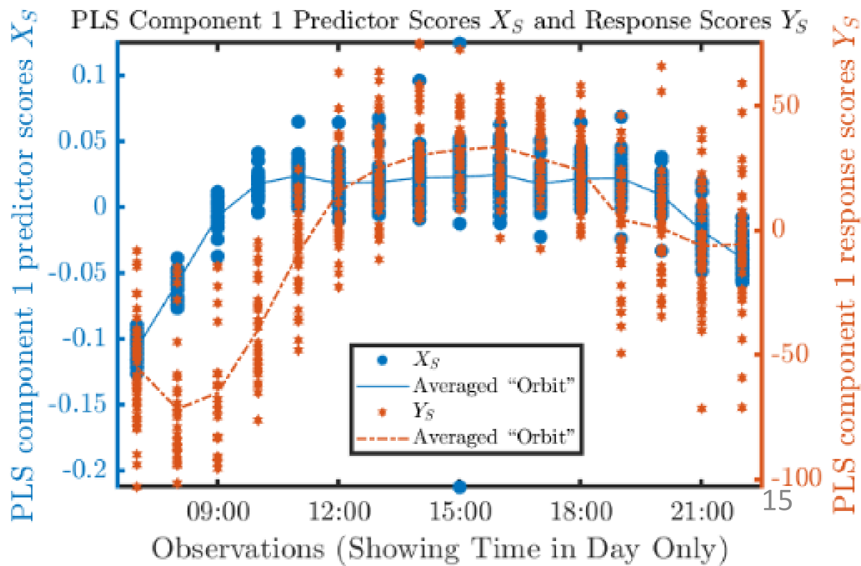


Interpretations and Insight from PLS Regression Modeling

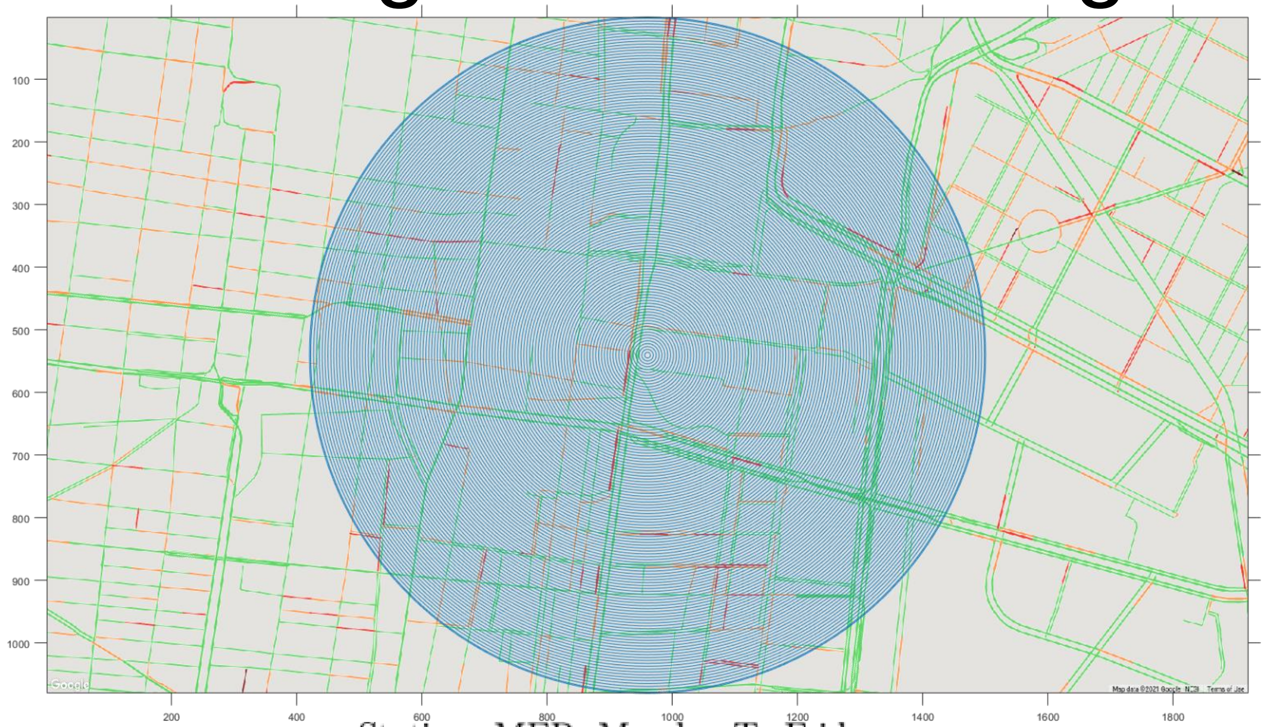


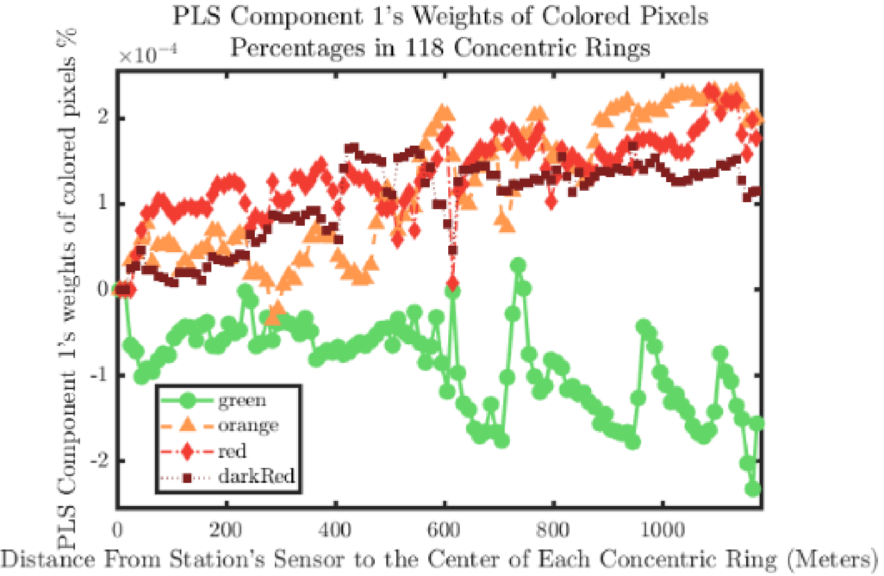


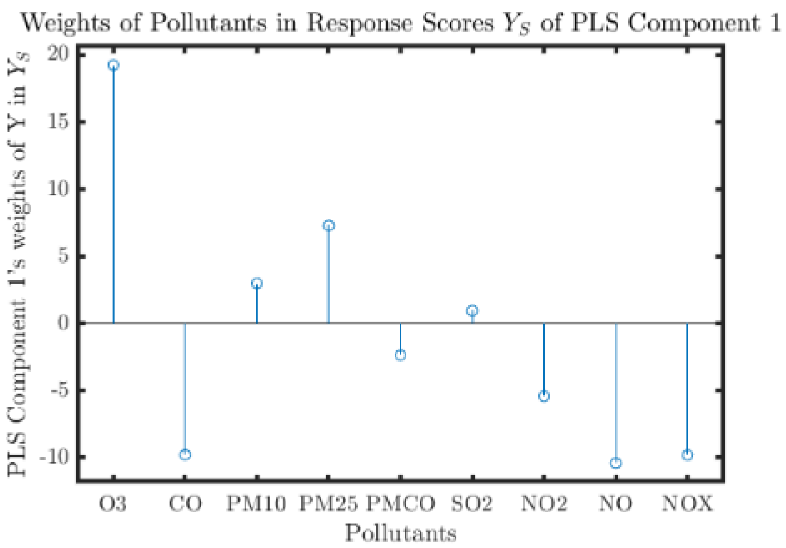


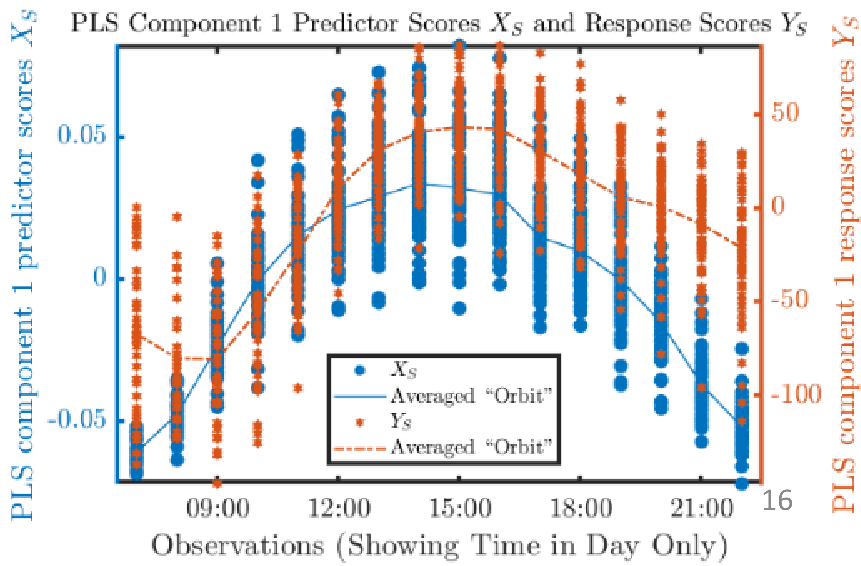


Interpretations and Insight from PLS Regression Modeling


Station: MER, Monday To Friday







Partial Lease Squares (PLS) regression and PCA regression

- When # of response variables in Y is small, we can use PCA to reduce the spatial dimension of X, and then fit a least squares model (Principal component regression, PCR)
- When responses Y is multi-dimensional/high dimensional, PLS regression often performs better.
- In actual model fitting, a (p predictors)-by-(m responses) coefficients matrix $\beta_{n_{\text{comp}}}$ is fitted in least squares sense for

$$Y_S Y_L^{\mathsf{T}} = X_S X_L^{\mathsf{T}} \beta_{n_{\text{comp}}}$$

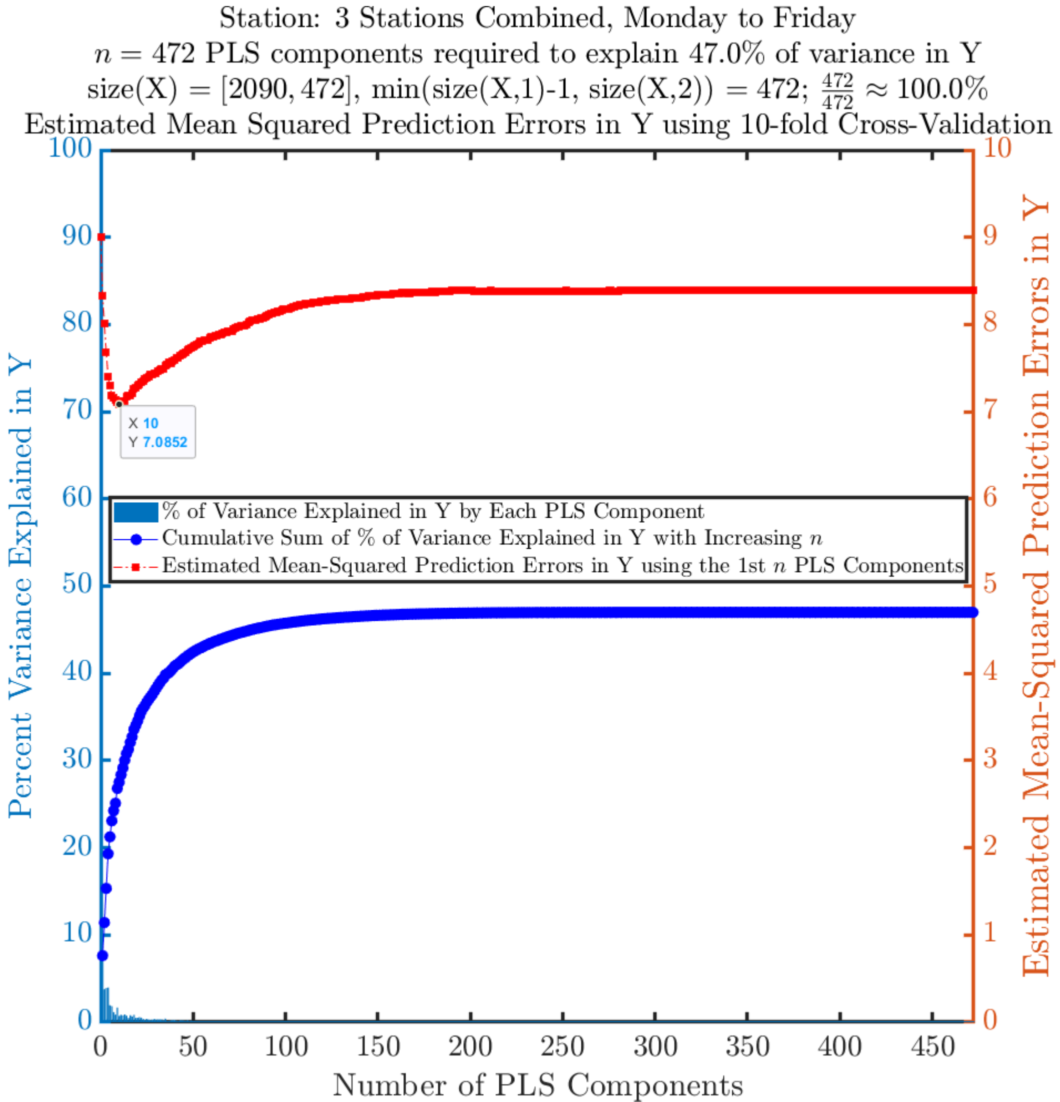
using truncated n_{comp} PLS components, and $X_S X_L^{\mathsf{T}}$ and $Y_S Y_L^{\mathsf{T}}$ are “reconstruction” of X_0 and Y_0

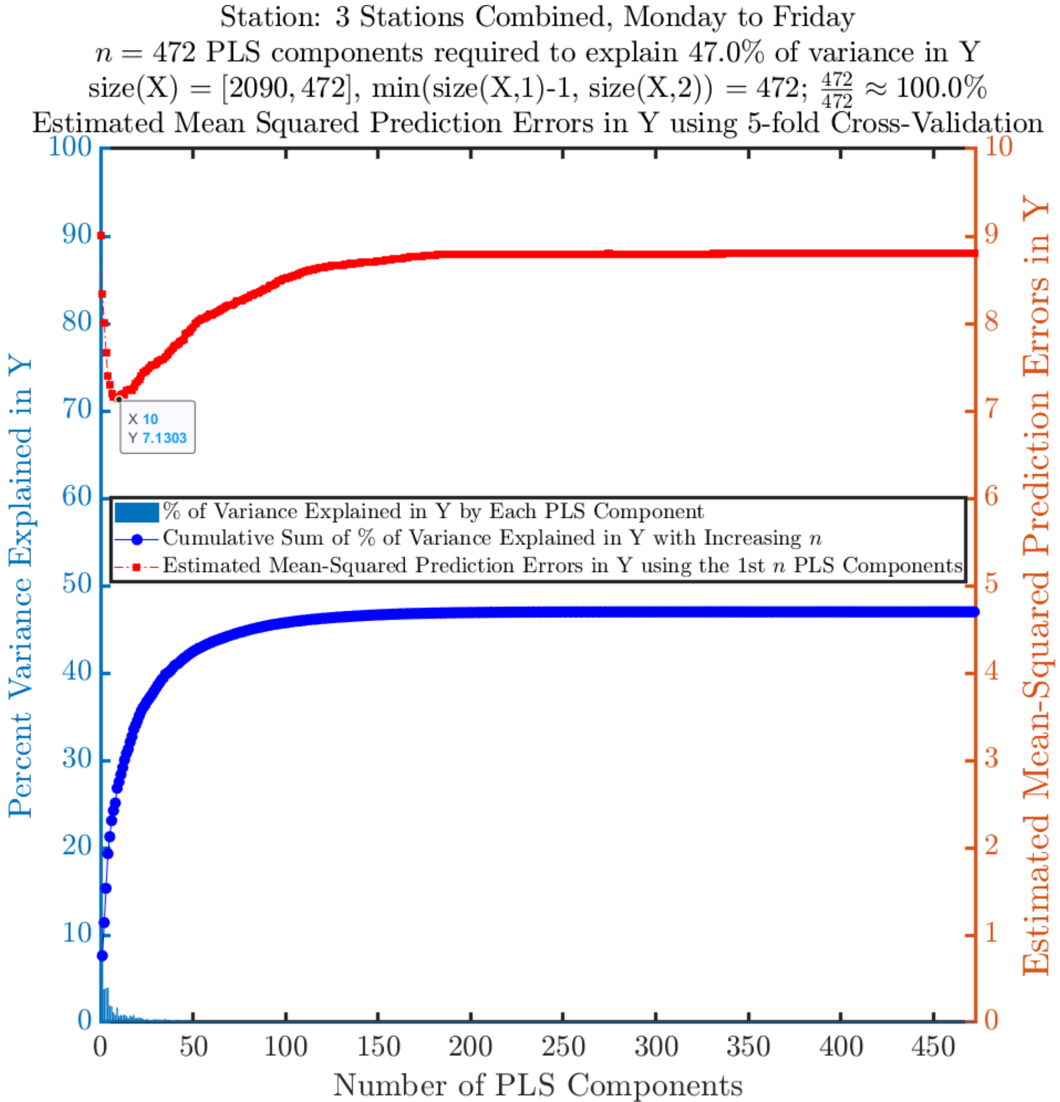
- Spatial dimension in X_0 reduced from p to n_{comp}, effectively fitting $Y_S = X_S \beta$ (a “partial” least-square), as compared to naïvely fit $Y = X\beta$
- n_{comp} can be fixed by cross-validation to minimize the expected mean-squared errors (MSE) $\left(Y_0 - Y_S Y_L^{\mathsf{T}}\right)^2$.
- (a preliminary result on next page)

PLSR Modeling and Prediction Performance

SAPIENS
Smart Air Pollution Information Enabling New Solutions

(all 9 pollutant response variables are centered to 0 mean with rescaled Var=1)





Outlook and Conclusions

(conclusions by Marcella)

SAPIENS has built a database with both pollution measurements and traffic images, so we have:

- Cleaned and analysed the data and identified patterns
- Developed a model to extract the traffic intensities from Google Map images
- Used the regression modeling to (1) obtain interpretable insights on the relation between traffic and pollutants; and (2) train it on the data from three stations (traffic and pollution data) and cross-validated it to avoid overfitting

On-going activities:

- Validation/testing phase: use other sensors data to validate/test model
- Paper in preparation

Outlook and Conclusions

There are more ideas and more possibilities to exploit and learn from these data.

More ideas on how to exploit the predicting power of the modeling

E.g., incorporating meteorological data, going beyond linear modeling techniques, etc.

Stay tuned for more from us

SAPIENS

SMART AIR POLLUTION INFORMATION ENABLING NEW SOLUTIONS